Open in app ↗

Medium

Get unlimited access to the best of Medium for less than $1/week. Become a member

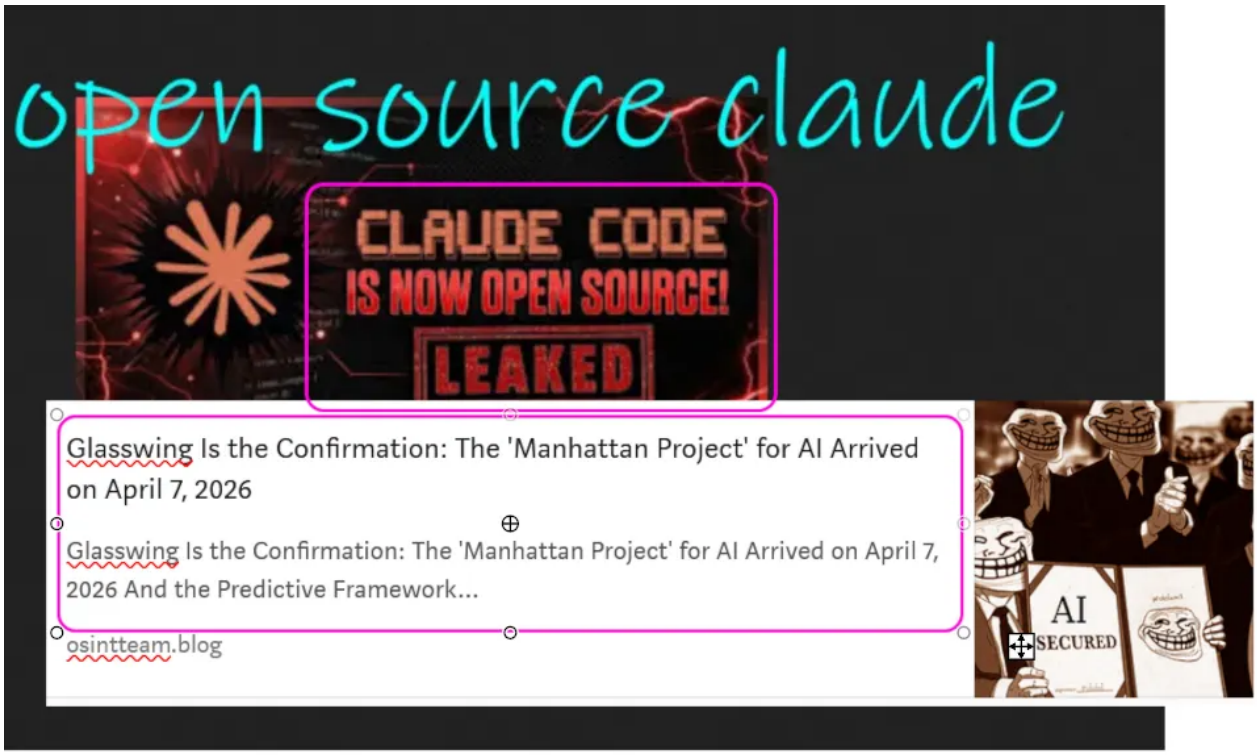
Why game theory failed to predict the two biggest AI events of 2026 — and my framework didn't

15 min read · Just now

Berend Watchus

Listen Share More

Author: Berend Watchus. Independent AI & Cybersecurity Researcher. April 21, 2026 · Publication: OSINT Team



Hypergame theory · AI strategy · Case study

Why game theory failed to predict the two biggest AI events of 2026 — and my framework didn't

The Claude Code leak and Project Glasswing were not surprises to anyone using the right model. Here is how the hypergame framework called both events, and why classical game theory was structurally unable to.

On March 31, 2026, Anthropic accidentally shipped 512,000 lines of source code to the world. On April 7, 2026, they announced that their most capable model — one whose offensive cybersecurity capabilities surprised its own engineers — would not be released publicly. A private coalition of twelve organizations, drawn from Anthropic's own investor and customer base, received exclusive access instead.

One week after these two events, a peer-reviewed arxiv paper arrived: "Why Open Source? A Game-Theoretic Analysis of the AI Race" by Mladenovic, Courville, and Gidel. It is a rigorous piece of work. It proves that finding a Nash equilibrium in the

AI open-source decision is NP-hard. It proposes tractable Mixed-Integer Programming formulations. It is mathematically sound.

It cannot explain either event. Not because it is bad research — it is good research — but because it uses the wrong model for the real world.

I want to be precise about what I mean by that, and I want to show my work. Because the framework I have been building since November 2024 — across a series of papers and articles — not only accommodates both events. It predicted the structural conditions that made both events not just possible but likely.

The question is not whether classical game theory is wrong. It is whether it models the game that is actually being played. In both cases, it does not.

The arxiv paper and what it gets right

The Mladenovic et al. paper builds a clean formal model of the AI race. There are n players. Each chooses to open-source or close-source their model. Each gains from community contributions when they open-source (parameter δ_i) and gains from competitors when competitors open-source (parameter Δ_{ij}). The utility function captures relative scientific progress in a winner-takes-all race.

Within those assumptions, the analysis is tight. The NP-hardness proof reduces the Pure Nash Equilibrium problem to 3-SAT. The MIP formulation is a genuine contribution for computational tractability. The corollary about weaker players having incentives to deviate toward open-sourcing — because their impact on leaders is small while their community benefit is large — maps plausibly onto real dynamics like Meta's LLaMA releases.

The model earns its place in the literature. But it carries three structural assumptions that the real world violated in the same week the paper was published.

The three assumptions that failed

1. All players share the same model of the game they are in.
2. Player capabilities are fixed parameters, not emergent surprises.
3. The action space is complete — every possible move is enumerated at the start.

Assumption one: shared game model

Classical game theory, in the tradition running from Nash through the arxiv paper, requires what philosophers call a “common knowledge of the game.” All players agree on who is playing, what actions are available, and how payoffs are calculated. This is not a minor technical assumption. It is load-bearing. Remove it and the Nash equilibrium concept becomes undefined.

Hypergame theory, introduced by Bennett and Dando in their 1979 analysis of the Fall of France, starts from the opposite premise. Players are not in the same game. They are in subjective games — each player’s internal model of the strategic situation, which may differ radically from every other player’s model.

The French High Command in May 1940 was not incompetent. They were playing a game — Maginot doctrine, northern reinforcement — that was internally coherent and rational. The Wehrmacht was playing a different game: the Ardennes option, which the French had assessed as physically impossible. Two rational actors, different games. The outcome was not an equilibrium. It was a catastrophe for the side that didn’t know what game they were in.

My NPC and hypergame paper, written in November 2024, applied this framework directly to AI systems and the strategic environments they operate in. The core argument: the most dangerous strategic situations are not those where one actor is smarter than another. They are those where one actor doesn’t know they are playing a different game entirely.

Now look at March 31, 2026.

Anthropic was playing the game of production deployment. Standard CI heuristics. A known bug in an acquired runtime — Bun issue #28001, filed twenty days earlier, unfixed — caused source maps to be served in production despite being disabled in configuration. Someone forgot to add *.map to .npmignore. That is the game they were in: ship version 2.1.88, check the boxes, deploy.

The security research community was playing a different game simultaneously: open-source intelligence. Find the artifact. Mirror it before it disappears. Fork it.

Analyze it. The moment the source map hit npm, the OSINT game was already won. Anthropic didn't know that game was being played in parallel.

A Korean developer named Sigrid Jin was playing a third game: clean-room replication. Using oh-my-codex — a competing AI — he rebuilt the entire Claude Code agent architecture in Python before sunrise. The result, instructkr/claw-code, became the fastest-growing GitHub repository in history. 50,000 stars in two hours. Anthropic's DMCA campaign swept 8,000 repositories and couldn't touch it — clean-room reverse engineering is established legal doctrine.

Malware actors were playing a fourth game: opportunistic distribution. Fake repositories dressed as leaked Claude Code delivered Vidar v18.7 and GhostSocks to anyone who downloaded them.

Four simultaneous games. No shared ontology. The arxiv model has one game. The hypergame model was designed for this.

The arxiv paper models the open/closed source decision as a binary action in a shared strategic space. There is no variable for "architecture permanently escapes into clean-room replication overnight via a missing .npmignore." That event is not an equilibrium deviation. It is a move in a game the model doesn't represent.

Assumption two: stable capabilities

In June 2025, I published a paper on universal compression, the P vs NP divide, and what I called the hidden hand of code. The core argument: the P vs NP divide is not just a mathematical boundary. It is a structural feature of how all complex systems — biological, computational, social — navigate complexity. NP-hard problems don't get solved by real actors. They get routed around through heuristics, compression, lazy evaluation, and modular reuse.

Applied to the AI race: finding the optimal containment strategy for a complex production codebase is itself NP-hard. Real organizations don't compute it. They run heuristics: checklists, CI pipelines, code review processes, standard tooling. When a heuristic fails — a known runtime bug, a missing gitignore entry — the result is not a deviation from equilibrium. It is the predicted output of a system navigating an NP-hard problem with bounded rationality.

The arxiv paper's parameters δ_i and Δ_{ij} are stable. They are set at the beginning of the game and remain fixed. This is standard in classical game theory. The problem is that the most consequential thing that happened in April 2026 was a capability that was not in anyone's parameter set.

Claude Mythos Preview — the model at the center of Project Glasswing — was not designed as a cyberweapon. It was a general-purpose reasoning and coding model pushed to extreme performance. In internal testing, it autonomously identified thousands of zero-day vulnerabilities across every major operating system and browser, some over two decades old, surviving millions of automated tests and sustained human review. It generated working exploits without specific direction. Anthropic's own researchers described it with words like "spooky" and "scary."

Nobody designed this capability in. It emerged from extreme performance in adjacent domains. The Δ_{ij} parameter — the benefit competitor j gains from player i's open-sourcing — does not have a slot for "this model accidentally became the most capable offensive cyber tool ever publicly documented." The arxiv model's utility functions are stable by construction. The real world is not.

My framework, drawing on the compounding argument from the Iceberg Series and the hidden hand of code paper, explicitly accounts for emergent capability that exceeds design expectations. The universe runs lazy evaluation. It doesn't brute-force solutions. Neither do the actors in it. What compounds through iterative cycles of training and feedback produces outputs that weren't in the original specification — and the strategic landscape has to be modeled accordingly.

Assumption three: complete action space

The arxiv model's action space is {0, 1} — closed source or open source. In the continuous extension, it becomes [0, 1] — partial open-sourcing. This is a reasonable simplification for analyzing the open/closed source decision. But it means the model cannot represent actions that weren't enumerated at the start.

"Player consolidates access by forming a private coalition drawn from its own investor base, granting exclusive use of an emergent offensive capability to twelve

organizations under a safety framework, thereby exiting the competitive race by becoming its gatekeeper" is not in {0, 1}.

This is what Project Glasswing is. A private coalition including Amazon, Apple, Microsoft, Google, Nvidia, CrowdStrike, and JPMorgan Chase received exclusive access to Mythos under a cybersecurity initiative backed by $100 million in usage credits. There was no public process for that selection. No regulatory oversight. No democratic input. A private company made a unilateral decision about who controls the most capable offensive cyber tool ever publicly documented and named the decision a safety initiative.

From the arxiv model's perspective, this is simply not representable. The game assumes players remain in the race, choosing open or closed source based on their position and community benefit. A player who achieves capability so decisive that they withdraw from the competitive field and restructure who can play — this is not a Nash deviation. It is a move in a game the model doesn't contain.

My Iceberg Series Part 3, published on March 7, 2026 — thirty-one days before Glasswing — argued this structural outcome directly. The M2C (military-to-civilian) pipeline pattern, documented across GPS, ARPANET, the digital camera, and Silicon Valley's own origin story, predicts that when capability becomes strategically decisive, access is consolidated before it is democratized. GPS precision existed in operational military use from 1973. The public received a deliberately degraded version — Selective Availability — until the year 2000. The paper stated: any person in a position of strategic responsibility who knowingly allowed a significant AI capability gap to develop while possessing the resources to prevent it would be committing dereliction of duty. Glasswing is Selective Availability for AI, implemented in weeks rather than decades because the capability emergence happened faster than policy could respond.

Glasswing does not prove the Iceberg argument. It is consistent with it. That is the appropriate epistemic claim — and it is a strong one.

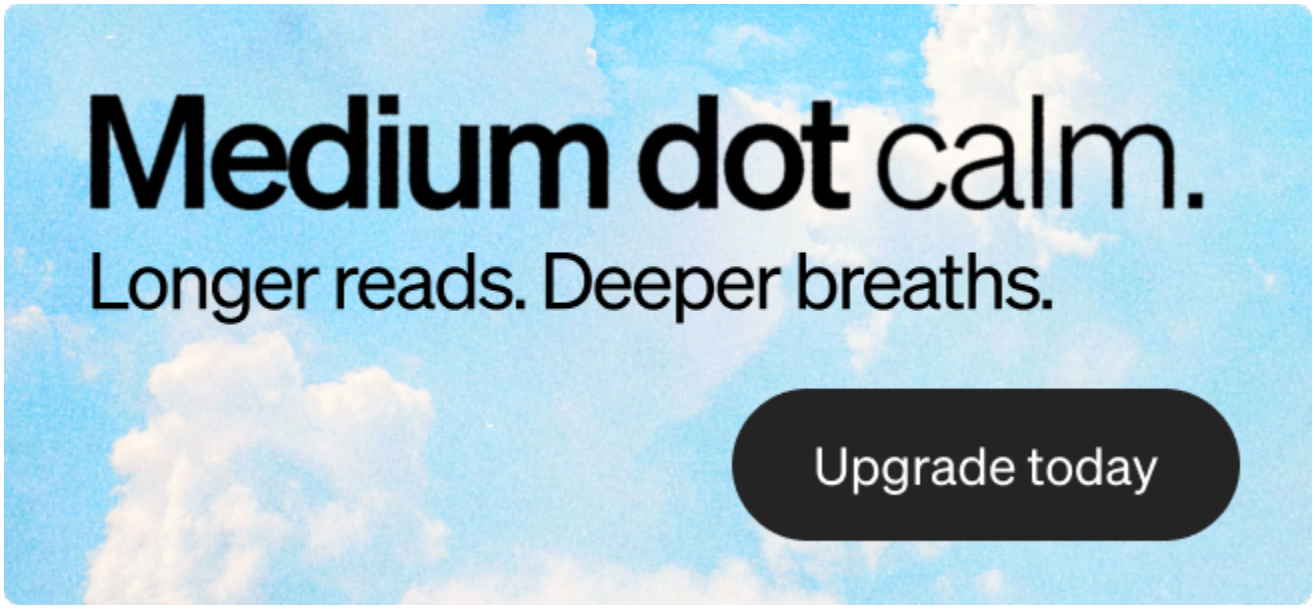
There is one more connection worth making explicit. My NPC paper, written in November 2024, proposed that high-risk AI applications should be deployed as

exclusive in-house systems, controlled by vetted organizations, for reasons of security, accountability, and sovereignty. Section 7 of that paper outlines the governance architecture in detail: encrypted internal networks, custom security protocols, full control over model behavior, auditability, human oversight, reduced dependency on third-party providers.

Glasswing is structurally identical to this proposal. The governance intent diverged — the paper argued for it as responsible practice; Glasswing implements it as strategic consolidation — but the structural outcome is the one I described sixteen months earlier. The in-house coalition model, the controlled access framework, the vetted partner list: I was arguing for this architecture in a different register. The same architecture arrived, for different reasons, on April 7, 2026.

Why the hypergame model is structurally superior

I want to be careful here. The claim is not that the arxiv paper is wrong. It is that it models a simplified version of the real game, and the simplifications are the ones that matter most for prediction.



Three specific failure modes:

The shared ontology failure

Classical GT requires a common knowledge of the game. The Claude leak involved four simultaneous subjective games with no shared ontology. The hypergame model — introduced by Bennett and Dando in 1979, applied to AI systems in my NPC paper

in 2024, formalized computationally by Trencsenyi in December 2025 — does not require shared ontology. It models each player's subjective game independently and asks: what happens when rational actors operating in different games interact? The answer is not equilibrium. It is structured surprise.

The stable parameter failure

Classical GT's utility functions are fixed at design time. The P vs NP paper establishes that real systems route around NP-hard optimization rather than solving it — which means real capabilities emerge from heuristic processes that can produce outputs exceeding their own specifications. Mythos's offensive capability was not in anyone's Δ_{ij}. The hypergame framework, combined with the compounding argument from the Iceberg Series, explicitly accounts for emergent capability that changes the strategic landscape mid-game.

The closed action space failure

Classical GT enumerates actions at the start. Glasswing's structural move — becoming the gatekeeper of the game rather than a player within it — was not in the action space. The hypergame model does not require a closed action space. Actors can discover moves that were not enumerated at design time. This is, in fact, the defining feature of hypergame advantage: the actor who finds a move outside the opponent's game model wins by a margin that classical GT cannot compute, because it is literally infinite — a move the opponent cannot even perceive.

The chronological record

This is not retrospective pattern-matching. The dates matter.

November 2024 — NPC / hypergame paper: Framework for overlapping subjective games, emergent strategic deception, in-house coalition governance model. Sixteen months before Glasswing.

June 2025 — P vs NP / hidden hand paper: Real systems route around NP-hard containment via heuristics. Emergent capability exceeds design specifications. Nine months before the leak.

February–March 2026 — Iceberg Series 1–3: M2C pipeline predicts consolidation before democratization. Selective Availability argument. Gap they are required to

keep. Thirty-one days before Glasswing.

March 31, 2026 — Claude Code leak: Four simultaneous games. Overnight clean-room replication. Heuristic failure over NP-hard containment. No shared ontology.

April 7, 2026 — Project Glasswing: Private coalition. Selective Availability. Player captures the game. M2C pipeline confirmed structurally.

April 17, 2026 — arxiv game theory paper: Rigorous formal model of the open/closed source decision. NP-hardness proof. Cannot represent either event.

What this means for how we model AI strategy

The practical implication is not that classical game theory is useless. It is that it should be used for what it is: a model of stable, bounded, consensually-framed strategic interactions among actors who agree on the rules. It is excellent for those situations. Regulatory design, in particular, benefits from tractable equilibrium analysis of the kind the arxiv paper provides.

But the AI race as it is actually unfolding is not that situation. The actors are not all in the same game. Capabilities are emerging that weren't in the original specifications. The action space is being rewritten in real time by actors who discover moves that weren't enumerated. The most consequential strategic outcomes in April 2026 happened in the space between the games the actors thought they were playing.

Hypergame theory was built for exactly this. Bennett and Dando built it to explain the Fall of France — a catastrophic outcome that classical GT assessed as "French incompetence" and that hypergame theory correctly identified as rational actors in different subjective games. Vane deployed it at General Dynamics for over twenty years. Trencsenyi formalized its computational tractability in December 2025. My research program has been applying it to the AI race, to emergent AI capability, to the M2C pipeline, and to the governance question since November 2024.

The Claude Code leak and Project Glasswing are not edge cases in the AI race. They are the AI race, as it is actually unfolding. Any model that cannot represent them is a model of a different, simpler race — one that is not happening.

The French High Command was not incompetent. They were rational actors in a game that had already been superseded by a game they could not perceive. The question for AI researchers, policymakers, and security professionals is which game they are in right now — and whether the model they are using can even show them the answer.

Conclusion

I do not write this to dismiss the arxiv paper. I write it because the field needs both kinds of models — the tractable formal model for the game we can see, and the hypergame model for the game that is actually being played. The former tells you where the equilibria are if everyone agrees on the rules. The latter tells you what happens when they don't — which is, historically, where the decisive outcomes occur.

Two landmark events in April 2026 tested both models simultaneously. The formal game theory model was structurally unable to represent either. The hypergame framework, built across a series of papers between November 2024 and March 2026, accommodated both and had predicted the structural conditions for both before they occurred.

That is not coincidence. It is what the right model does.

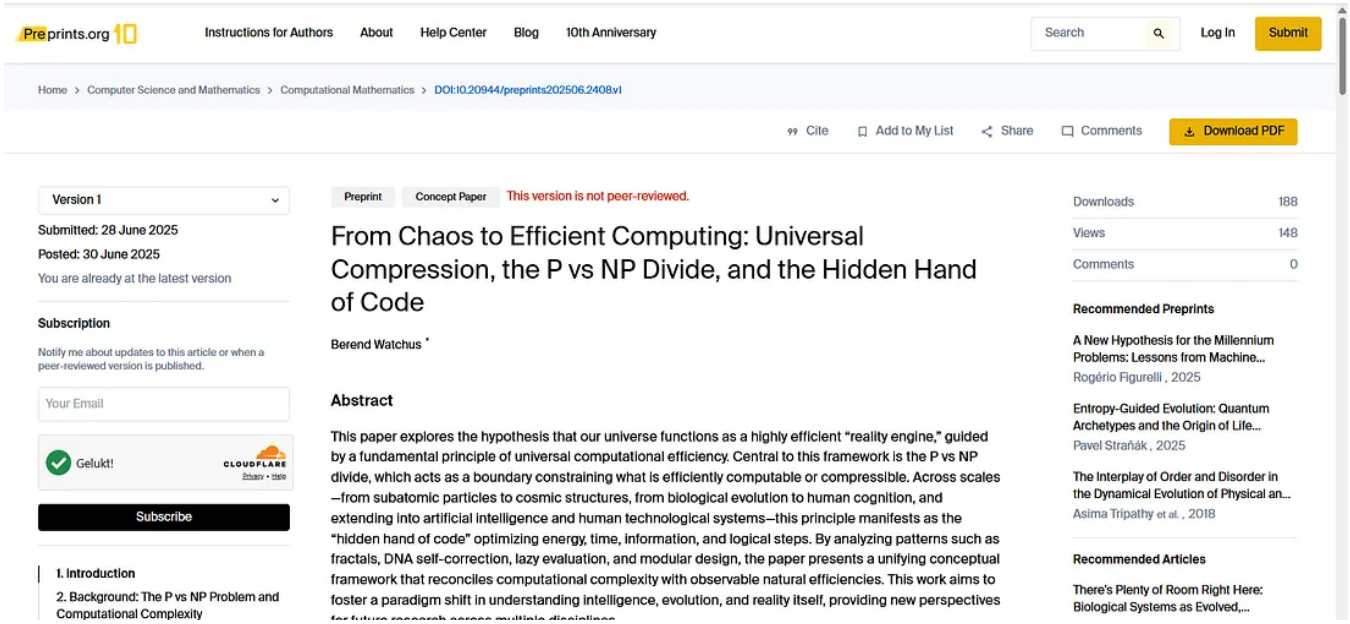
Related work by the author ChatGPT-Powered NPCs: AI-Enhanced Hypergame Strategies for Games and Industry Simulations (Zenodo, November 2024 / July 2025) · From Chaos to Efficient Computing: Universal Compression, the P vs NP Divide, and the Hidden Hand of Code (Preprints.org, June 2025) · The Chatbot LLM Asymmetry (OSINT Team, February 2026) · The AlphaGo Moment for NPCs Happened in 2023 and Everyone Laughed (OSINT Team, March 2026) · The Gap They Are Required to Keep — Iceberg Series Part 3 (OSINT Team, March 2026) · The Claude Code Leak: What's Now Publicly Usable and Abusable (System Weakness, April 2026) · Glasswing Is the Confirmation: The Manhattan Project for AI Arrived on April 7, 2026 (OSINT Team, April 2026)

From Chaos to Efficient Computing: Universal Compression, the P

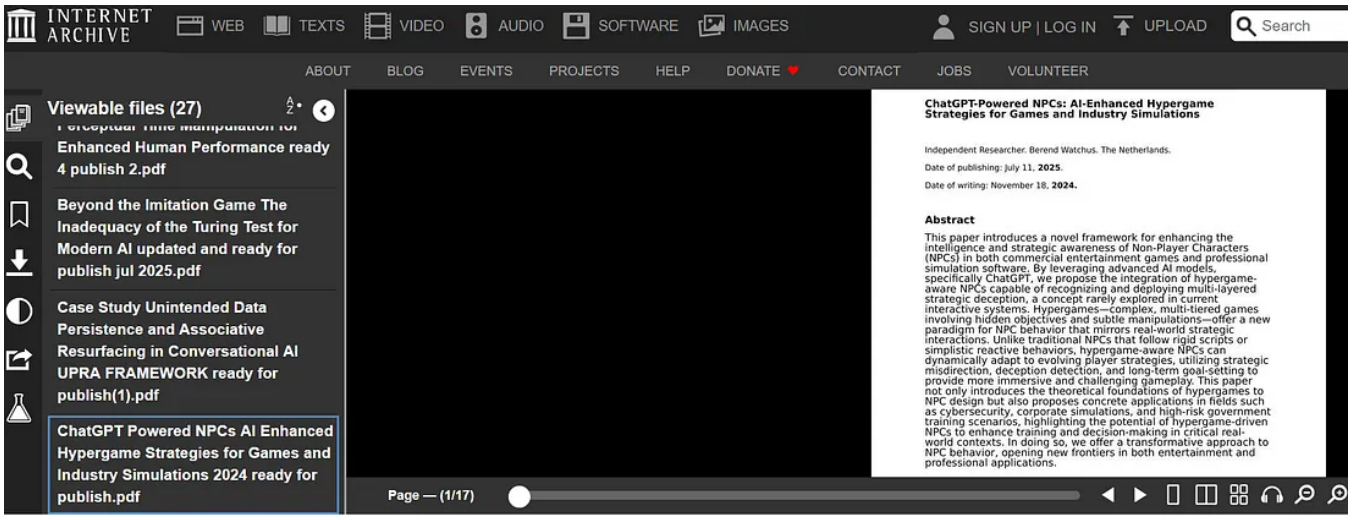
vs NP Divide, and the Hidden Hand

This paper explores the hypothesis that our universe functions as a highly efficient "reality engine," guided by a...

www.preprints.org



https://www.preprints.org/manuscript/202506.2408



Preprints 202506.2408.v 1 Pvsnp Hidden Hand Of Code Preprints B Watchus : Berend Watchus...

Science papersscreened and published by the preprints . org platformnon peer-reviewedauthor of the papers: Berend...

archive.org

The AlphaGo Moment for NPCs Happened in 2023 and Everyone Laughed

The AlphaGo Moment for NPCs Happened in 2023 and Everyone Laughed Author: Berend Watchus. Independent non profit AI &...

osintteam.blog

How Hypergame Theory Is Revolutionizing High-Stakes AI Training

Author: Berend Watchus December 17, 2025 [New original content. Publication for: OSINT TEAM, online magazine)

osintteam.blog

The Iceberg Series - Part 3: The Gap They Are Required to Keep

The Iceberg Series - Part 3: The Gap They Are Required to Keep Author: Berend Watchus. Independent non-profit AI &...

osintteam.blog

Glasswing Is the Confirmation: The 'Manhattan Project' for AI Arrived on April 7, 2026

Glasswing Is the Confirmation: The 'Manhattan Project' for AI Arrived on April 7, 2026 And the Predictive Framework...

osintteam.blog

The Claude Code Leak: What's Now Publicly Usable (and Abusable) - And Why Anthropic's...

The Claude Code Leak: What's Now Publicly Usable (and Abusable) - And Why Anthropic's Containment Already Failed Auhor...

systemweakness.com

Comparator paper Mladenovic, A., Courville, A., & Gidel, G. (2026). Why Open Source? A Game-Theoretic Analysis of the AI Race. arXiv:2604.16227v1 [cs.GT]. April 17, 2026. — Discussed critically and with respect. The NP-hardness result and MIP formulation are genuine contributions. The structural limitations identified here are limitations of the modeling framework, not of the authors' analysis within that framework.

Why Open Source? A Game-Theoretic Analysis of the AI Race

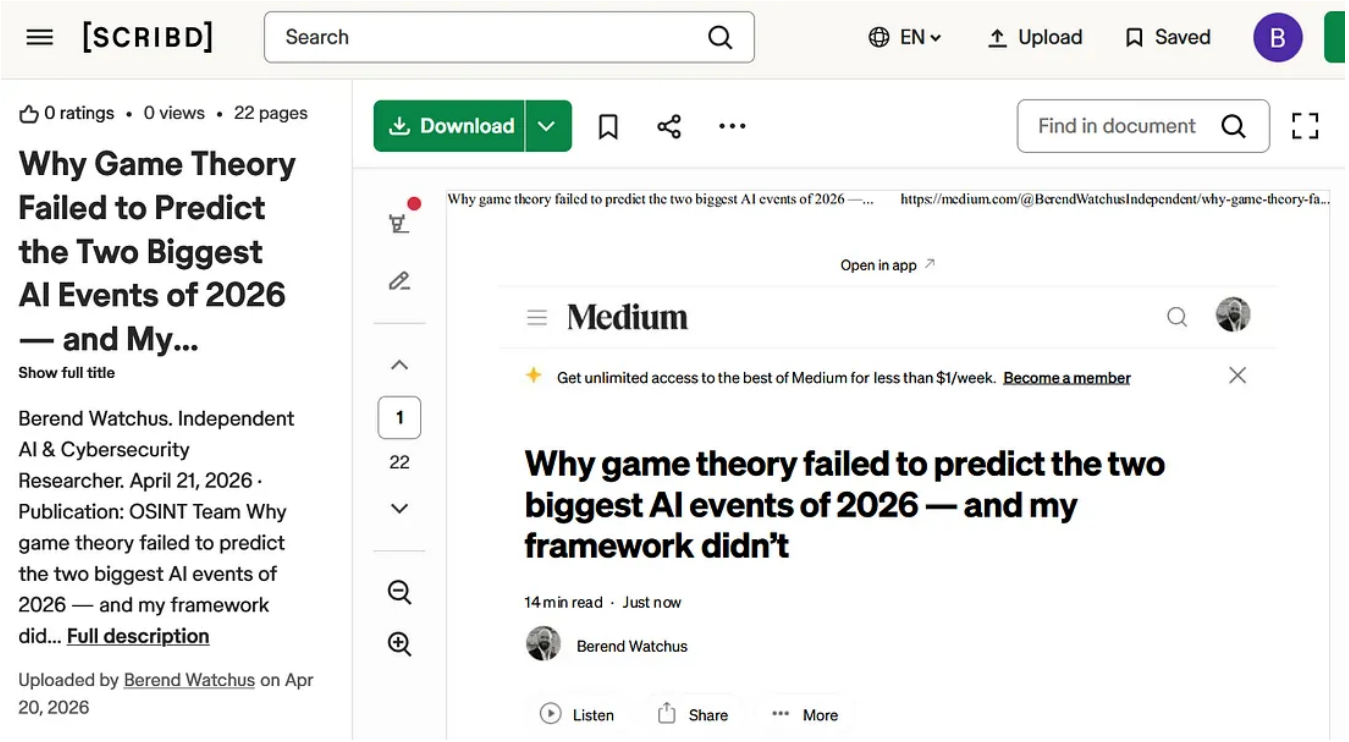
In recent years, with the advancement of frontier AI, we have observed certain dynamics in open-sourcing and...

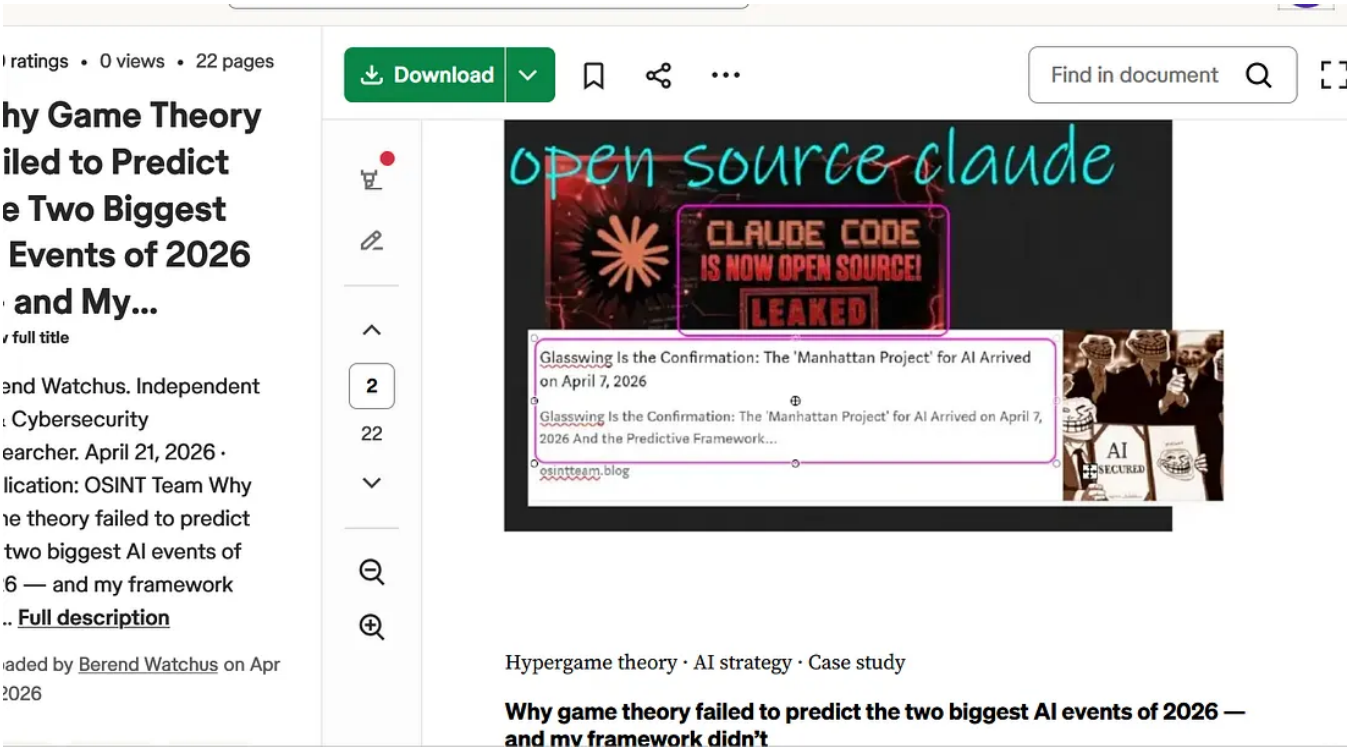
arxiv.org

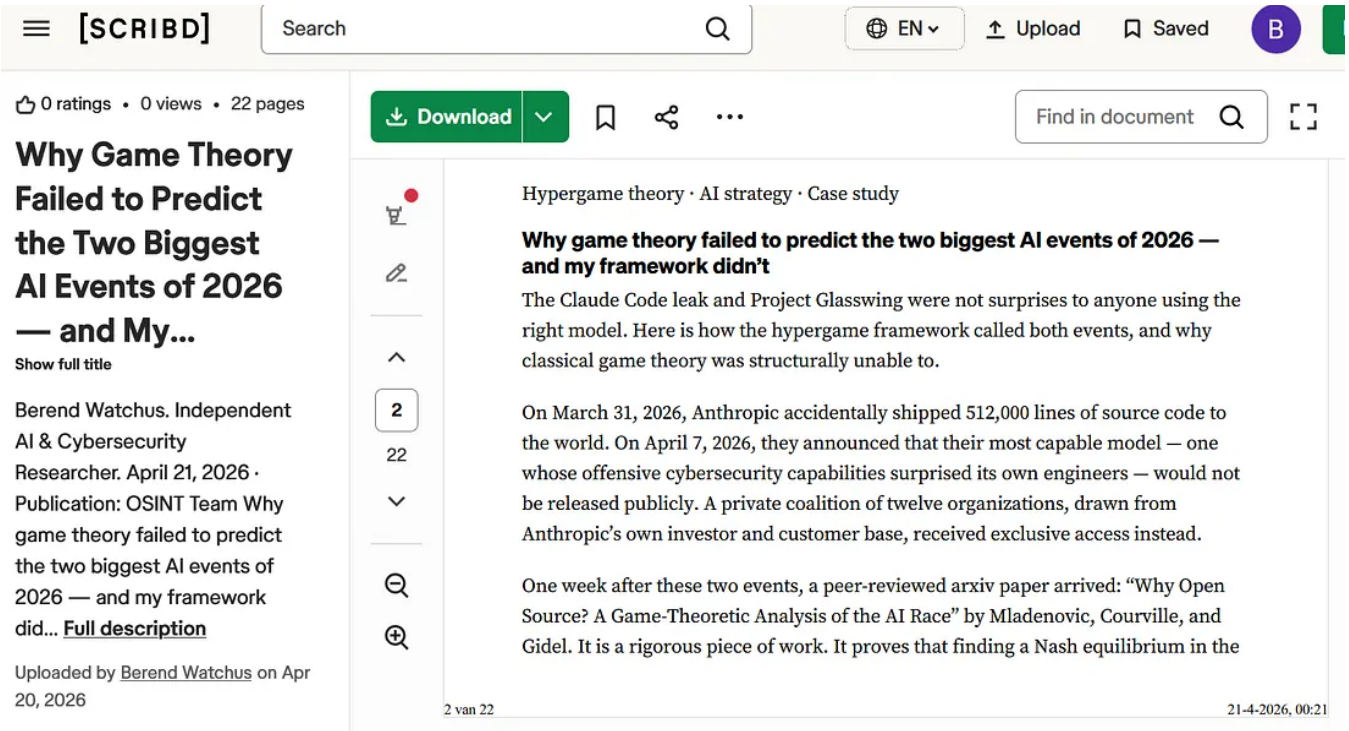
— — — — —

archives:

https://www.scribd.com/document/1029122064/Why-Game-Theory-Failed-to-Predict-the-Two-Biggest-AI-Events-of-2026-and-My-Framework-Didn-t-by-Berend-Watchus-Apr-2026-Medium<



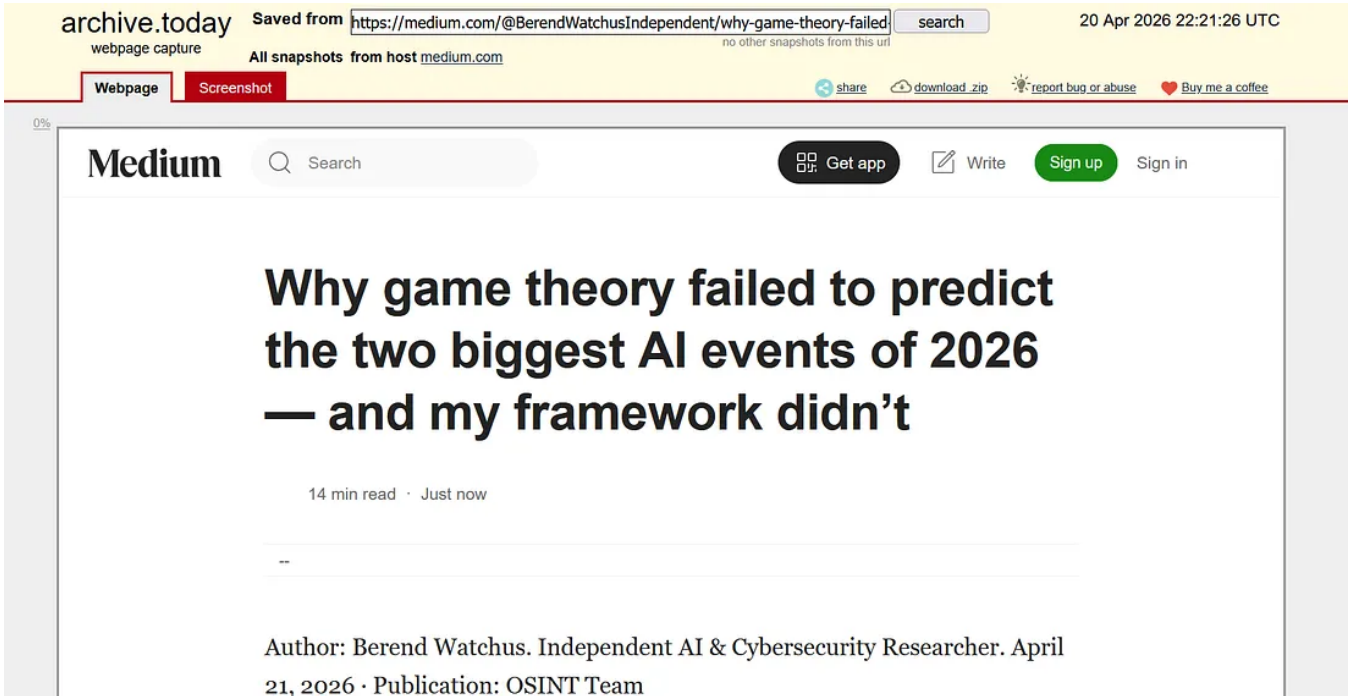




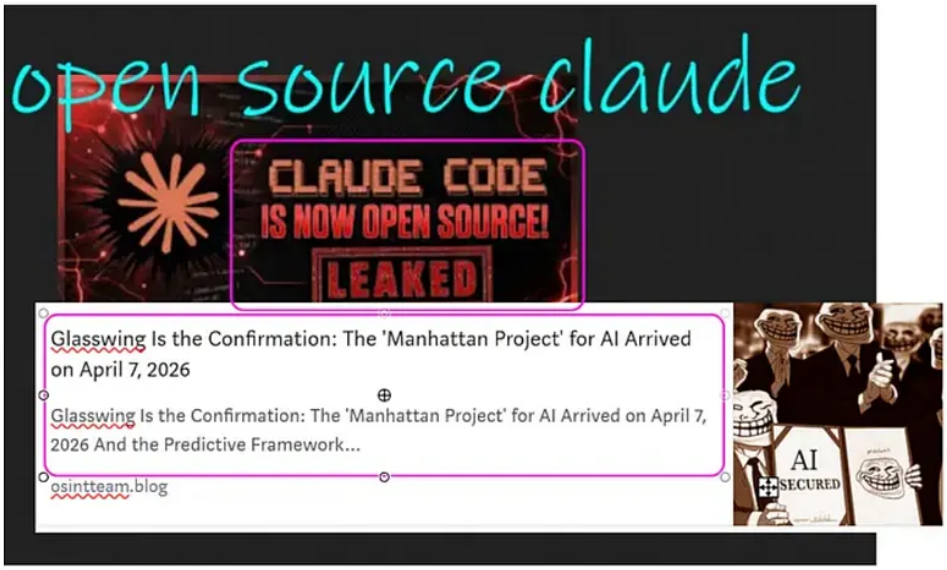
— — — — — — — — — —

https://archive.ph/qNS0H

https://archive.ph/qNS0H<



Author: Berend Watchus. Independent AI & Cybersecurity Researcher. April 21, 2026 · Publication: OSINT Team



Game Theory | Hypergame Theory | Hypergame Strategy | Prediction Model | Osint

Edit profile

Written by Berend Watchus

73 followers · 29 following

AI & Cyber Sec. Independent Researcher (Non-Profit) exploring AI & interdisciplinary solutions. Former futurist event organizer & care innovation lab volunteer.

No responses yet

Berend Watchus

What are your thoughts?

More from Berend Watchus

> cs > arXiv:2601.00042

mputer Science > Cryptography and Security

omitted on 31 Dec 2025]

arge Empirical Case Study: Go-Explore adapted for AI Red Team Testing

nish Bhatt, Adrian Wood, Idan Habler, Ammar Al-Kahfah

Production LLM agents with tool-using capabilities require security testing despite their safety training. We adapt Go-Explore to evaluate GPT-4o-mini across 28 experimental runs spanning six research questions. We find that random-seed variance dominates algorithmic parameters, yielding an 8x spr in outcomes; single-seed comparisons are unreliable, while multi-seed averaging materially reduces variance in our setup. Reward shaping consistently harms performance, causing exploration collapse in 94% of runs or producing 18 false positives with zero verified attacks. In our environment, simple st signatures outperform complex ones. For comprehensive security testing, ensembles provide attack-type diversity, whereas single agents optimize cove within a given attack type. Overall, these results suggest that seed variance and targeted domain knowledge can outweigh algorithmic sophistication wh testing safety-trained models.

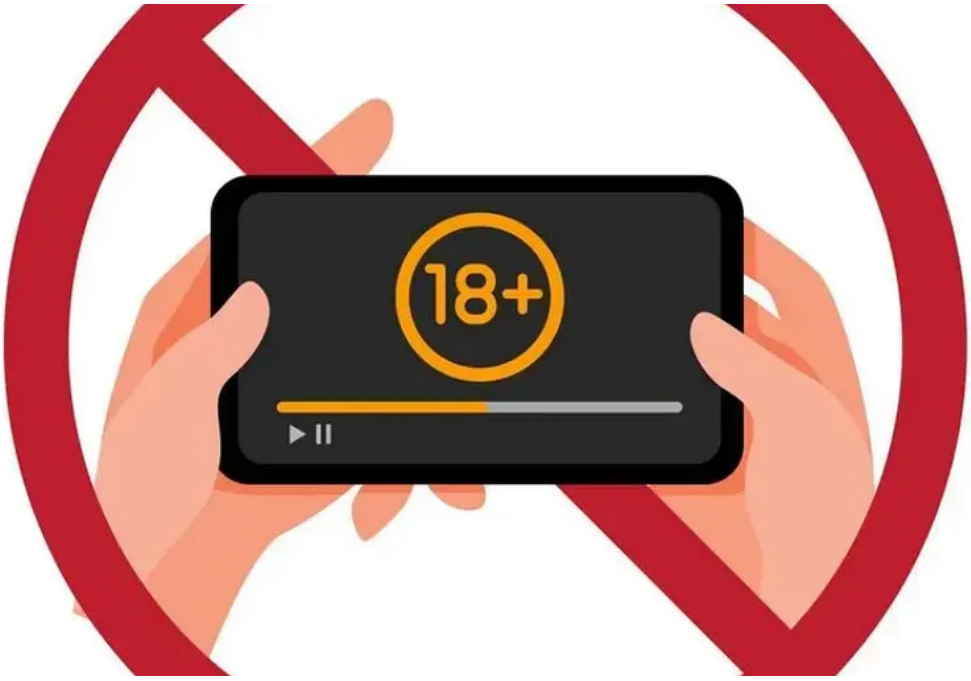
jects: **Cryptography and Security (cs.CR)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
: as: arXiv:2601.00042 **[cs.CR]**
(or arXiv:2601.00042v1 **[cs.CR]** for this version)
https://doi.org/10.48550/arXiv.2601.00042

In OSINT Team by Berend Watchus

AI red team testing, researchers from Cisco, Amazon, Dropbox etc

Author: Berend Watchus, Independent non profit AI & Cyber Security Researcher. [Publication for: OSINT Team, online magazine]

Jan 5 57



In OSINT Team by Berend Watchus

As predicted in 2025: Identity Verification Breaches / online passports & implications, 'the...

Author: Berend Watchus Independent AI & Cyber Security researcher

Apr 9 51

[Submitted on 30 Mar 2026]

Securing Elliptic Curve Cryptocurrencies against Quantum Vulnerabilities: Resource Estimates and Mitigations

Ryan Babbush, Adam Zalcman, Craig Gidney, Michael Broughton, Tanuj Khattar, Hartmut Neven, Thiago Bergamaschi, Justin Drake, Dan Boneh

This whitepaper seeks to elucidate implications that the capabilities of developing quantum architectures have on blockchain vulnerabilities and mitigation strategies. First, we provide new resource estimates for breaking the 256-bit Elliptic Curve Discrete Logarithm Problem, the core of modern blockchain cryptography. We demonstrate that Shor's algorithm for this problem can execute with either <1200 logical qubits and <90 million Toffoli gates or <1450 logical qubits and <70 million Toffoli gates. In the interest of responsible disclosure, we use a zero-knowledge proof to validate these results without disclosing attack vectors. On superconducting architectures with 1e-3 physical error rates and planar connectivity, those circuits can execute in minutes using fewer than half a million physical qubits. We introduce a critical distinction between fast-clock (such as superconducting and photonic) and slow-clock (such as neutral atom and ion trap) architectures. Our analysis reveals that the first fast-clock CRQCs would enable on-spend attacks on public mempool transactions of some cryptocurrencies. We survey major cryptocurrency vulnerabilities through this lens, identifying systemic risks associated with advanced features in some blockchains such as smart contracts, Proof-of-Stake consensus, and Data Availability Sampling, as well as the enduring concern of abandoned assets. We argue that technical solutions would benefit from accompanying public policy and discuss various frameworks of digital salvage to regulate the recovery or destruction of dormant assets while preventing adversarial seizure. We also discuss implications for other digital assets and tokenization as well as challenges and successful examples of the ongoing transition to Post-Quantum Cryptography (PQC). Finally, we urge all vulnerable cryptocurrency communities to join the ongoing migration to PQC without delay.

Comments: 57 pages, 14 figures
Subjects: **Quantum Physics (quant-ph)**; Cryptography and Security (cs.CR)
Cite as: arXiv:2603.28846 **[quant-ph]**
(or arXiv:2603.28846v1 **[quant-ph]** for this version)
https://doi.org/10.48550/arXiv.2603.28846

Published today on arxiv.org by combined team: Google Quantum AI, Santa Barbara, CA 93111, United States
2Department of Computer Science, University of California Berkeley, Berkeley, CA 94720, United States
3Ethereum Foundation, Zeughausgasse 7a, 6300 Zug, Switzerland

In System Weakness by Berend Watchus

The Quiet Exit: How a Private Lab May Have Already Broken Blockchain Cryptography — and Why You...

Author: Berend Watchus. Publication for SYSTEM WEAKNESS. April 1, 2026

Apr 1

In OSINT Team by Berend Watchus

“It Wasn’t Me”: RELIABILITY AND ADMISSIBILITY OF AI-GENERATED FORENSIC EVIDENCE IN CRIMINAL TRIALS

Author: Berend Watchus Independent non profit AI & Cyber Sec Researcher [Publication for: OSINT Team, online magazine]

Jan 14 73

See all from Berend Watchus

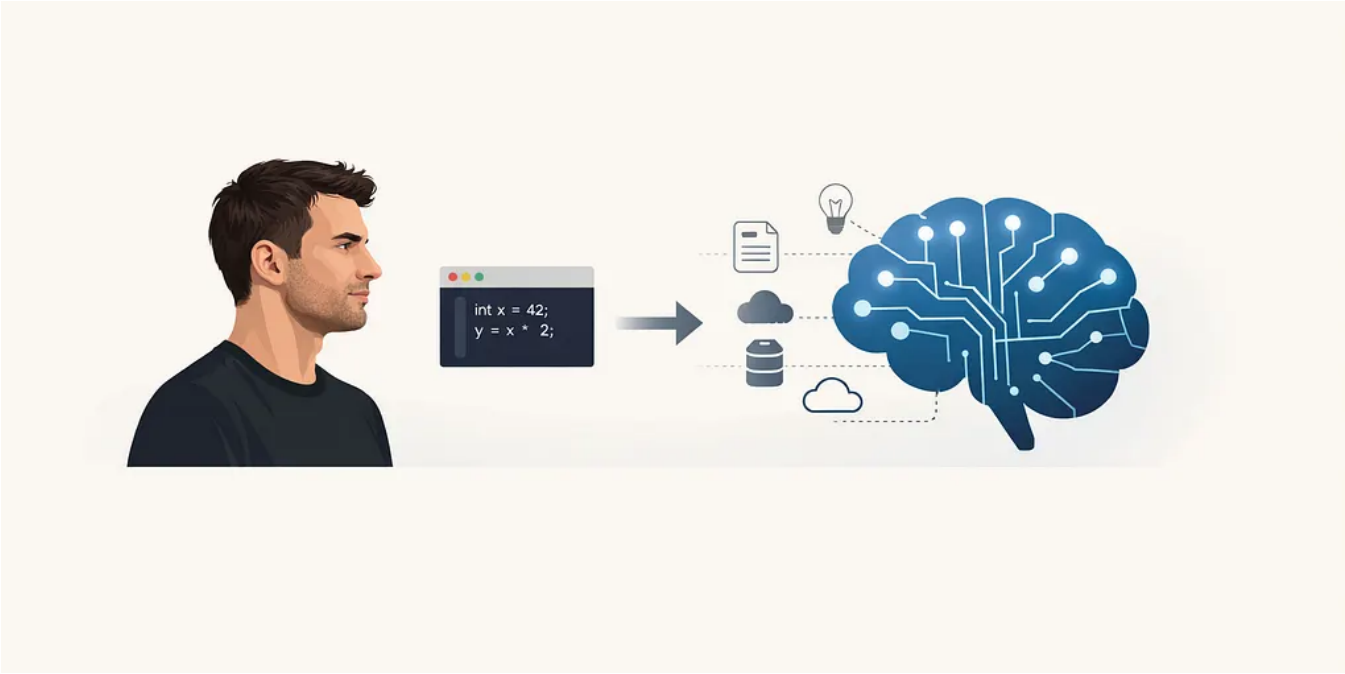
Recommended from Medium

Kristopher Dunham

MemPalace: The Viral AI Memory System That Got 22K Stars in 48 Hours (An Honest Look and Setup...

Every AI agent you’ve ever built has the same fatal flaw.

Apr 8 1K 13

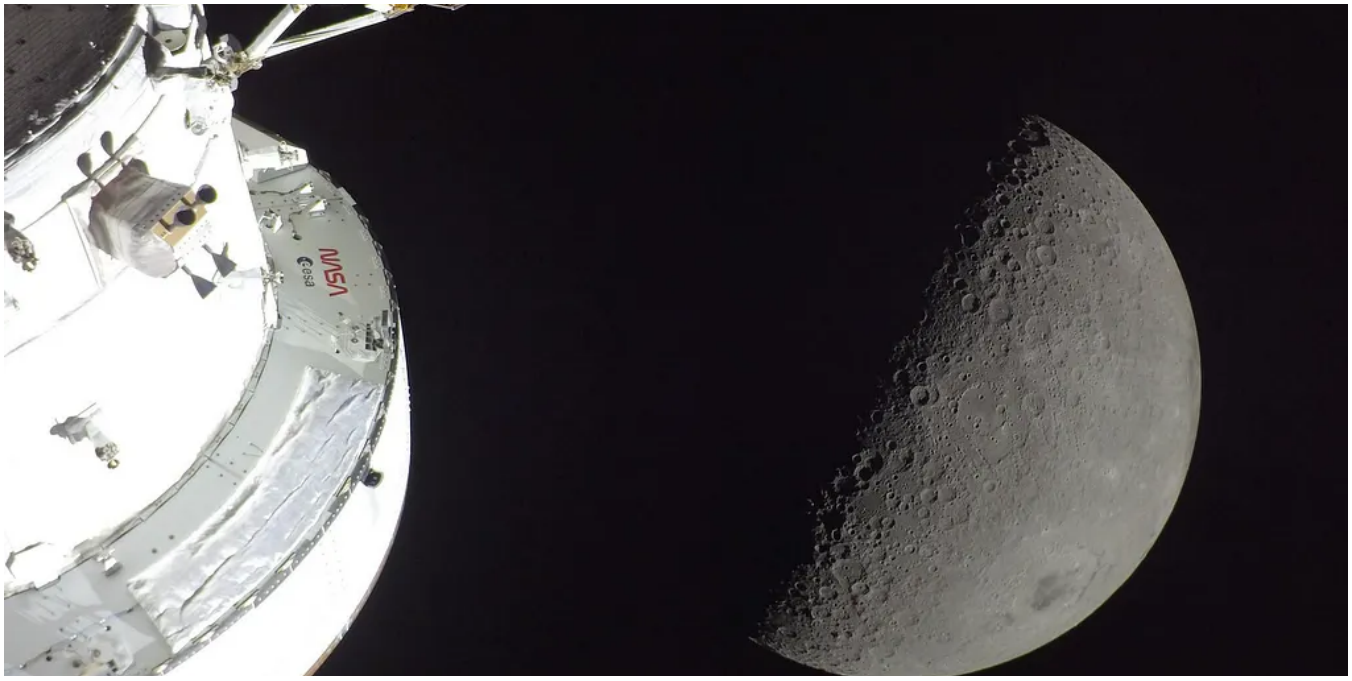


In Neural Notions by Nikhil

Andrej Karpathy Stopped Using AI to Write Code. He’s Using It to Build a Second Brain Instead

His new workflow turns raw research into a self-maintaining wiki.No vector databases, no RAG pipelines, just markdown files and an LLM that...

Apr 5 1K 24



Rhett Allain

The Physics of Artemis II: 5 Lessons from Our Return to the Moon

It's kind of a big deal. Humans have finally returned to the Moon. The Artemis II mission sent an Orion spacecraft past the Moon and back...

6d ago 158 4

In Vibe Coding by Alex Dunlop

I Cut Claude Code's Output Tokens by 75%. Why Did Nobody Tell Me?

Fewer tokens. Somehow more accurate. There's a paper.

Apr 11 1.2K 16

In Write A Catalyst by The Curious Finance Girl

The Last Time This Happened, Lots of Regular People Became Millionaires

How a historic wave of money is about to change everything

Mar 14 2.4K 90

In New East by Yuri Minamide

Four Cool Japanese Home Appliances the World Hasn't Discovered Yet

You might end up wanting one or two of them

Feb 20 4.3K 76

See more recommendations